Processing IonTorrent Sequencing Data using NextGENe Software

December 2010

John McGuigan, Megan Manion, Kevin LeVan, CS Jonathan Liu

Introduction

The new IonPGM™ Personal Genome Machine made by Ion Torrent a part of Life Technologies enables fast and inexpensive next-generation sequencing using real-time measurement of hydrogen ions released during DNA replication. NextGENe now includes a module specially designed for processing data from the PGM™. Data from amplicon or small genome sequencing projects can be easily and rapidly analyzed using a point-and-click windows interface on a standard desktop computer. In most cases it takes less than 3 minutes to go from raw data to a mutation report.

Procedure

Project setup is guided in a point-and-click interface by NextGENe's project wizard. The instrument type and analysis options are first selected as seen in figure 1. NextGENe's Format Conversion Tool is used to convert data from SFF or FASTQ format to FASTA format. Filtering and trimming based on quality scores is performed at the same time. Suggested settings are shown in figure 2, but due to the rapidly improving quality of PGM™ sequence data, different quality settings may provide more optimal results. Next, the data, reference, and output location are specified as seen in figure 3. Finally the alignment options including mutation filter settings are chosen as seen in figure 4.

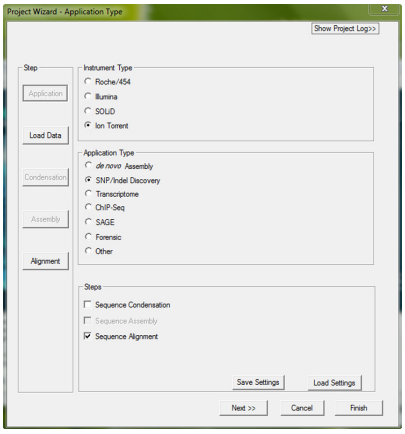


Figure 1

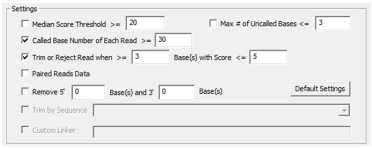


Figure 2

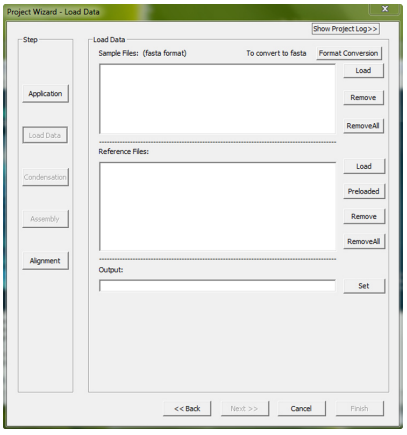


Figure 3

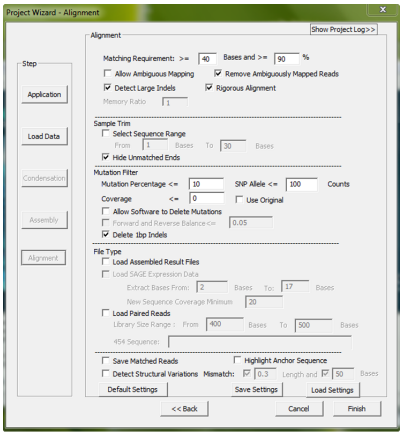


Figure 4

Multiple runs can be set up to run sequentially or they can be set up and run one at a time. When the project has completed running it is opened in the NextGENe viewer. The highly customizable mutation report lists all of the variants that passed the mutation filter. Multiple other reports are available for review, including an expression report to determine coverage levels for each amplicon whose locations can be specified with a BED file as seen in figure 5.

Index	Contig	Chr	Chr Position St	Chr Position End	Gene	CDS	Length	Max Counts	Average Cou	Read Counts	Forward Read	RPKM
1	NG_007572	1	862168	862228	NRAS; +	2	61	11543	11193.00	11543	6361	2061660.4913
2	NG_023302	15	726002	726061	IDH2; +	4	60	9859	9533.70	9859	5549	1790234.4247
3	NG_007572	1	859957	860013	NRAS; +	1	57	9105	8643.50	9105	5002	1740337.1151
4	NG_007726	7	353007	353146	EGFR; +	20	140	12579	6884.10	12597	5159	980318.9130
5	NG_007524	12	837077	837160	KRAS; +	2	84	5715	5583.20	5715	3003	741250.9047
6	NG_007873	7	171384	171454	BRAF; +	15	71	5626	5515.90	5626	2774	863315.7555
7	NG_007666	11	691734	691801	HRAS; +	2	68	4890	4706.20	4890	3013	783480.5764
8	NG_007726	7	346432	346561	EGFR; +	19	130	6230	4326.10	6280	4801	526313.5840
9	NG_023319	2	700056	700145	IDH1; +	2	90	3886	3747.50	3886	1482	470423.0297
10	NG_007524	12	838779	838863	KRAS; +	3	85	3231	3141.90	3231	942	414139.1808
11	NG_007726	7	363491	363618	EGFR; +	21	128	3550	2520.40	3553	2987	302422.1006
12	NG_007524	12	819069	819137	KRAS; +	1	69	2541	2467.50	2541	1354	401221.1904
13	NG_007456	4	805936	806033	KIT; +	17	98	2205	2114.00	2205	966	245138.0945
14	NG_007726	7	345376	345766	EGFR; +	18	391	7677	1743.60	7677	2443	213915.9024
15	NG_007666	11	691317	691397	HRAS; +	1	81	1541	1508.50	1541	1166	207274.5150
16	NG_023302	15	726098	726173	IDH2; +	4	76	1536	1477.40	1536	921	220194.2182

Figure 5

Results

Figure 6 shows a 15 bp deletion detected in an amplicon resequencing project when aligning to a fasta file.

Figure 7 show the alignment of bacterial genome data using annotated gbk files as the reference, which allows annotation to be shown in the viewer. A synonymous mutation in the gadB gene is visible, highlighted in blue. Consensus and reference nucleotide and amino acid sequences are shown, as are gene (blue arrow) and coding sequence (gold arrow) locations. The gray shading indicates depth of coverage. Figure 8 shows the distribution report of the aligned reads- an even distribution of reads in the forward and reverse direction is good for more accurate mutation calling. The reads average over 125 bp in length.

Figure 9 shows two point mutations- one homozygous and one heterozygous- detected in another amplicon resequencing project.

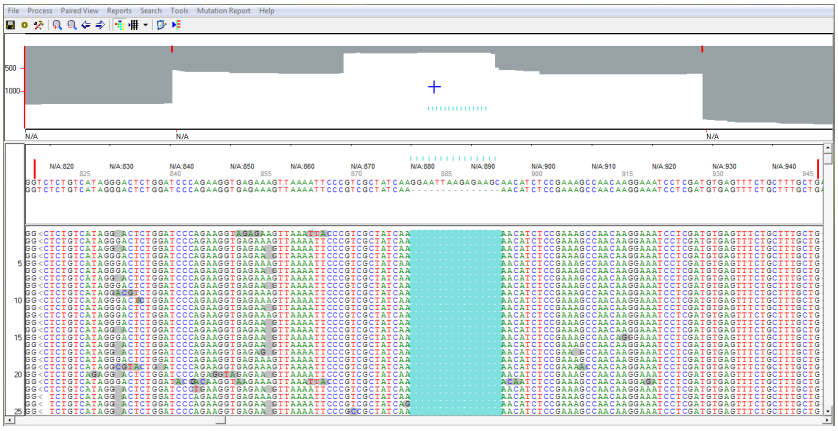
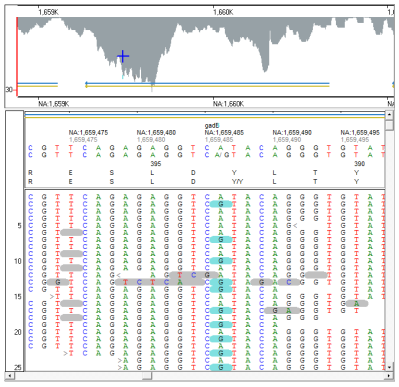


Figure 6



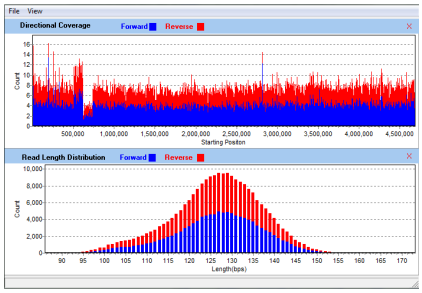


Figure 8

Figure 7

Discussion

NextGENe uses a specialized hash alignment method for processing IonTorrent data in order to account for its unique error profile. Additionally, small deletions in homopolymer regions are assumed to be errors and are highlighted in the viewer but not included in the mutation report. This option ("Delete 1bp Indels") can be disabled within the alignment settings.

The mutation filter settings should be adjusted based on the expected depth of coverage and frequency of mutations in order to optimize specificity and sensitivity. The minimum matching percentage can also be adjusted to match more total reads or to match only the reads that have few errors.

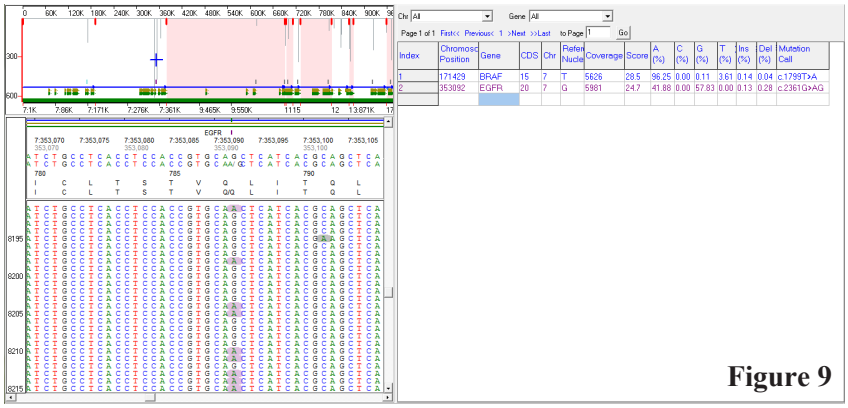


Figure 9

NextGENe provides a proprietary mutation confidence score for every called mutation. The maximum score is related to the coverage (8*log10(coverage)) and several penalty scores lower this score from that maximum. It may be convenient to disable the "Read Balance" and "Alelle Balance" penalty scores within the mutation report settings if the data is not expected to be directionally balanced.

Acknowledgements

Thanks to Dr. Long Phi Le, a molecular pathologist, and Dr. John Iafrate, Director of the Diagnostic Molecular Pathology Laboratory, both from Massachusetts General Hospital, for providing data.